Is Justice Really Blind? And Is It Also Deaf?

Daniel L. Chen[1]

Abstract

Using data from 1946–2014, I show that audio features of lawyers' introductory statements and lawyers' facial attributes improve the performance of the best prediction models of Supreme Court outcomes. I infer face attributes using the MIT-CBCL human-labeled face database and infer voice attributes using a 15-year sample of human-labeled Supreme Court advocate voices. I find that image features improved prediction of case outcomes from 64% to 69%, audio features improved prediction of case outcomes from 67% to 69%, and image and audio features together improved prediction of case outcomes from 64% to 67%. Lawyer traits receive approximately half the weight of the most important feature from the models without image or audio features. When it comes to predicting Justice votes as opposed to Supreme Court outcomes, image and audio features did not improve upon baseline models. This difference is consistent with human biases being more relevant in close cases.

[1] Daniel L. Chen, daniel.chen@iast.fr, Toulouse School of Economics, Institute for Advanced Study in Toulouse, University of Toulouse Capitole, Toulouse, France; dchen@law.harvard.edu, LWP, Harvard Law School. First draft: May 2016. Current draft: January 2018. Most recent version at: [http://users.nber.org/~dlchen/papers/Is_Justice_Really_Blind.pdf]. I thank Manoj Kumar, Vishal Motwani, and Phil Yeres for research assistance. Work on this project was conducted while I received financial support from the European Research Council (Grant No. 614708) and Agence Nationale de la Recherche.

Introduction

The emphasis on fit as a hiring criterion has raised the specter of a new form of subtle discrimination. Under complete markets, correlations between malleable characteristics and outcomes should not persist. Yet, using data on 1,901 U.S. Supreme Court oral arguments between 1998 and 2012, voice-based snap judgments based on lawyers' identical introductory sentences, “Mr. Chief Justice, (and) may it please the Court?”, predict court outcomes (Chen, et al. 2016, 2017). Understanding the causes (and consequences) of major court rulings has long been a topic of interest to social scientists and legal scholars (e.g., Rosenberg 1991).[2] At the same time, whether non-relevant factors such human voice or physical appearance predicts outcomes in high-stakes settings, as opposed to laboratory settings, has been the subject of much debate among scientists and psychologists (e.g., Todorov, et al. 2005). I bring these two concerns together in this paper.

A large body of work examines how people speak – their vowels, pitch, diction, and intonation – but there is relatively little evidence that speech variation beyond lexical choices (fluctuations in the way one speaks holding the words fixed) matters in real-world behavior, and this is where our paper comes in. Are vocal cues – and for that matter, visual cues – relevant in high-stakes policy-making settings such as the U.S. Supreme Court?

There are many reasons why physical impressions should not matter. From a rational perspective, information should override first impressions. From an ideological perspective, court outcomes are largely political and predetermined outcomes. From a legal perspective, legal decisions should be based on the legal content of the argument. From an economic perspective, correlations between malleable characteristics and outcomes should not persist as law firms and advocates adjust to eliminate such correlations.[3]

[2] Malleability of moral reasoning by judges has been documented in U.S. federal circuit judges (Ash, et al. 2016; Berdejo, et al. 2013; Chen, et al. 2017; Chen 2017), federal district judges (Chen 2017; Chen, et al. 2017), immigration judges (Chen, et al. 2016), sentencing judges (Chen 2017; Chen, et al. 2017), and military judges (Chen 2013). Some of these findings can be attributed to snap judgments or early predictability of judicial decisions based on race or nationality (Chen, et al. 2017; 2017).

[3] There are also a variety of institutional factors that can affect decisions on what individuals think is the fair and just (Chen et al. 2016; Shaw et al. 2011). Outside the lab, the malleability of injunctive norms to formal institutions such as the law (Chen et al. 2014; Chen et al. 2014; Chen et al. 2014) or markets (Chen 2015; Chen et al. 2014; Chen 2011) is suggestive of the impact of broader historical shifts in human rights (Chen 2004), sexual harassment (Chen et al. 2011), and free speech (Chen 2014).

I also build on a literature using limited samples of Supreme Court oral arguments that finds, for example, Supreme Court outcomes are correlated with authors' coding of emotional arousal in the behavior of Justices, lawyers, and their voices (Schubert, et al. 1992), are correlated with the number of questions asked by Justices (Epstein, et al. 2010), and are correlated with measurements of the emotional content of questions using linguistic dictionaries (Black, et al. 2011). These studies tend to have limited sets of covariates and employ linear regression models.[4]

Datasets

My Supreme Court data relies on case and court features for the 1946–2014 period. The data includes seven types of features: a) Justice and Court Background Information (e.g., Justice year of birth), b) Case Information (e.g., Legal Issue), c) Overall Historic Supreme Court Trends (e.g., Ideological Direction), d) Lower Court Trends (e.g., Circuit Court Ideological Trend), e) Current Supreme Court Trends (e.g., Mean Agreement Level of Current Court), f) Individual Supreme Court Justice Trends (e.g., Mean Justice Ideological Direction), and g) Differences in Trends (e.g., Difference Between Justice and Circuit Court Directions). My goal was to predict Supreme Court Justice votes and Supreme Court case outcomes. My data comes from Katz et al. (2014).

My labeled face data comprise 2,222 face images with ratings for 40 traits (e.g., confidence) from Bainbridge, et al. (2013). Each face was rated by 15 raters. I obtained roughly 1,000 faces of Supreme Court advocates representing 70% of the advocates who appeared before the Supreme Court over the 1971–2014 period.

My labeled voice data comprise 1,913 Supreme Court advocate audio clips from 1998–2012 with ratings for voice characteristics (confidence, masculinity, trust, intelligence, attractiveness and aggressiveness). Each audio clip was rated by approximately 20 Mechanical Turk workers, and a total of 20,888 ratings are available in this database. The data comes from Chen et al. (2016). I also have 14,932 unrated audio clips of Supreme Court advocates from 1946–1997 and 2013–2014.

All audio clips involve the lawyers' opening statements. These are identical sentences, "Mr. Chief Justice, (and) May It Please the Court", which means my analysis focuses on the way the advocate speaks while holding fixed the words they use.

Baseline Model and Performance Evaluation

[4] Their data are also not publicly available at the time of this writing.

My models build off of Katz, Martin and Bommarito's Supreme Court decision prediction model, available on Github.[5] Note that the October 2015 model available on their Github repository differs somewhat from the model described in their 2014 working paper. The data involved in both the models are the same, however, the 2014 model preprocessed[6] the data and tuned hyperparameters[7], while the 2015 model used a "growing ensemble" technique.[8]

Katz et. al. obtained the target labels from the Supreme Court Database (Epstein et. al), which codes for three possible outcomes:

1. 1: the justice reversed or the outcome in the lowercourt was overturned.
2. 0: the justice affirmed or the outcome in the lower court was upheld.
3. -1: the justice outcome was unable to be determined.

I removed the cases labeled as -1 to reduce the problem to a binary classification problem because it was my intuition that this was noise and predicting noise does not make much sense. Just by removing these labels and using the baseline model configuration, the justice-wise accuracy rose from 0.656 to 0.682 and the case-wise accuracy from 0.679 to 0.689. In all comparisons that follow (i.e., adding audio and image features), I focus on the binary classification problem for this reason.

Since the usual grid search with stratified split may be inappropriate because of the time-dimension of my data, I modified the baseline model with a custom time-series cross

[5] https://github.com/mjbommar/scotus-predict

[6] The original model scaled the features to have zero mean and unit variance. Then the most significant percentile features were selected on the basis of the ANOVA score. Then they trained an ExtraTreesClassifier on the preprocessed features.

[7] For hyperparameter tuning, the original model performed a grid search on the following hyperparameters: a) features selected in the preprocessing step, b) number of estimators of the extra trees classifier, c) minimum number of samples required in each leaf of the extra trees classifier, d) maximum depth of each tree of the extra trees classifier, and e) a subset of candidate features selected in each node of the tree. It used a 10-fold cross-validation strategy that preserves the percentage of classes in every split. The score used to identify the best hyperparameter on the test split was the F-score.

[8] A schematic of their code is as follows: Using as inputs–the minimum training period, minimum estimators, and trees per term–then a) Initialize train_period=min_train_years and estimators=min_estimators, b) Train the sklearn model on the data corresponding to the train_period, c) Predict on the year "train_period + 1", d) Increment "train_period" by 1 and estimators by "trees_per_term", and e) Repeat from 2, until the training period covers the last but one year. It is then trivial to calculate the mean accuracy score because I have the true and predicted values. All models set the training period to 5 years.

validation approach.[9] I searched randomly across parameters, and for every parameter sampled, I chose the parameter that gave the best mean accuracy across all years. I searched on maximum depth to which every tree is grown, number of features selected at every split, minimum number of samples in a leaf, and minimum number of samples in each node for it to be considered a split node. I sampled 30 random combinations of the parameters and found the best justice-wise and case-wise accuracy to be 0.674 and 0.666 respectively.[10]

Katz et al.'s model uses a large number of judge and case characteristic features, as well as court trend and lower court trend features. However, their model does not include advocate audio or image-based features. In the next sections, I describe how I generated audio and image-based features, and I demonstrate their effect on model performance.
It is important to note that as I add features to the model, I draw comparisons between the baseline model accuracy and the model incorporating the new features. In order to make the comparison apples-to-apples, I limit the data used to train the baseline model to the same cases where the new feature (image or audio) data is available. For instance, if I are comparing the model incorporating image features to the baseline model, I train the baseline model only on the cases where I have image ratings. It is for this reason that the baseline model accuracy varies in my model comparisons.

Features and Feature Engineering

Given the nature of the human face and voice, one design choice I encountered was whether to employ the underlying raw data on faces and voices or to use factors, such as trait features (for audio ratings from 1998–2012) or predicted trait features (for faces from 1971–2014 and for audio clips from 1946–1997 and 2013–2014). I chose to rely on predicted trait features rather than the underlying eigenvectors as features. This approach is commonly used in macroeconomic forecasting that relies on principal components or factor analysis. The underlying factor driving multiple economic indicators (eigenvectors) is believed to have continuous distribution. Moreover, since eigenvectors underlying common trait characteristics are likely to be highly correlated, a sparse model like LASSO is less appropriate. Both principal components analysis and regularization approaches aim to reduce dimensionality. However, regularization is a type of supervised learning

[9] This custom time-series cross-validation was inspired by a work-in-progress scikit-learn pull request. See "Add TimeSeriesCV and HomogeneousTimeSeriesCV" 2016. Github. Accessed July 4 2016. https://github.com/scikit-learn/scikit-learn/pull/6351.

[10] Since this randomized search approach decreased the case-wise and justice-wise accuracy in comparison to the default configuration, I assumed that tuning the hyperparameters of the random forest do not perform better than the default settings. Hence I use the default configuration throughout the rest of my analysis.

(considering the relationship between the outcome and predictors), whereas principal components analysis is a type of unsupervised learning (considering only the predictors), so using (predicted) trait features is more appropriate for my research question.

Image Features:

I trained models to predict facial trait ratings (confidence, unfriendliness, etc.), and I used those models to predict ratings for SCOTUS advocates. I then used the predicted ratings for SCOTUS advocates as features in the SCOTUS decision prediction model.

The human-labeled database of 2,222 images of faces comes from the MIT CBCL database.[11] The labels are ratings on a 0-9 scale for 40 traits (e.g., confidence, happiness, etc.). Motivated by Rojas, et. al. (2011), I performed the Histogram of Oriented Gradients (HOG) method on the images. This had the effect of converting the image to a sketch of the contours of the face. For every HOG-processed image, I vectorized the image's pixel matrix, converting it from a 100 x 128 matrix to a 1 x 12,800 vector. I took all the image vectors, stacked them into a matrix of dimension (# of images) X (12,800), and performed principal component analysis (PCA) on that matrix. I found that the top 100 principal components provided an explained variance ratio of 65%.

Using the top 100 principal components for each image as the features in 40 ridge models with inbuilt cross-validation (one for each trait), I built 40 trait rating prediction models.[12] I evaluated the 40 trait rating models and found that some have low mean squared error (MSE) and fairly high R^2. In fact, the HOG method substantially improved the MSE and R^2 of the ridge models. A full list of model statistics is available on request, which includes a comparison of MSEs and R^2s with and without HOG.

Next, I applied facial trait prediction models to attorney images. I collected images of lawyers for about 70% of the lawyers appearing before the Supreme Court during the 1971–2014 period. Then, I used the OpenCV Python package to locate faces in the images, and the Python Image Library to crop and resize the images to capture only the face. I then applied HOG to the images, vectorized the pixel matrix for each image, and performed PCA to transform the data into a matrix of image vectors. For each lawyer, I applied the 40 models trained for the traits and generated 40 predicted trait ratings.

[11] Image attribute data comes from the CBCL file, "psychology-attributes.txt".

[12] Because Bainbridge's image rating data does not contain rater identifiers, I could not normalize each rater's ratings.

Finally, because there can be multiple lawyers on the petitioner or the respondent side, I averaged ratings for lawyers on each side.

Audio features:

Next, I used a database of 1,913 audio clips representing SCOTUS advocates' opening remarks during oral arguments.[13] This data included clips in the 1998–2012 period, and, for each clip, there are associated voice trait ratings from humans. The ratings were on a 1-7 scale. I first normalized each rater's rating by subtracting their average rating and dividing by the standard deviation of their ratings (i.e. z-score). I then aggregated the z-scores corresponding to every lawyer thus giving me continuous voice trait ratings for every lawyer. I then made the z-scores binary: if a z-score was positive, we replaced it with 1, if it was negative I replaced it with -1.

Next, I processed every audio clip of lawyer's opening statements from 1946–2014 into fixed number of frames and I obtained the 13 Mel-frequency Cepstral Coefficients (MFCC)[14] for each of these frames. I vectorized the matrix of every audio clip, thus obtaining vectors of length 13, times the number of frames, for every audio clip from 1946–2014.

I trained two types of models to predict traits for audio clips from 1946–1997 and 2013–2014. The first type was trained prediction models using the continuous human rated audio clip data for the period 1998–2012. But I abandoned this approach due to low R^2 scores on test set. The second type was a trained random forest classifier model from the data in the period 1998–2012 using the binary score. I found that the second type model was successful at prediction and most accurate in predicting masculinity (65.79%) while least accurate in predicting trustworthiness (56.02%).[15] A full list of model stats is available on request. Finally, I applied the voice trait prediction model to lawyer voices in the period 1946–1997 and 2013–2014.

[13] This data was collected in Chen, Halberstam, and Yu (2017).

[14] Practical Cryptography "Mel Frequency Cepstral Coefficient (MFCC) tutorial". 2016. practicalcryptography.com. Accessed July 4 2016. Available at http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency- cepstral-coefficients-mfccs/. Lyons, James "Python Speech Features Repository". 2016. Github. Accessed July 4 2016. https://github.com/jameslyons/python_speech_features.

[15] The greater predictability in perceived masculinity is consistent with some results reported in Chen et. al. 2016, which plays the voice clips backwards and asks raters to rate the backward clips. Among the perceptual questions, ratings for perceived masculinity were most strongly correlated for the forward and backward clips.

These voice trait ratings were appended to the original data set. For the audio clips from 1998–2012, the binarized version of the originally obtained continuous z-score ratings were appended and for audio clips from 1946–1997 and 2013–2014, the binary voice traits predicted from the above mentioned model were appended.

Results

I find that image features improve the case-wise accuracy of the baseline model by 1.8 percentage points and decrease the justice-wise accuracy of the baseline model by 0.5 percentage points. Chen, et al. (2017) also report in a linear regression model that vocal features predicted case outcomes but not judge votes, and this is interpreted as being due to the swing voter being swayed by human biases in close cases. In brief, the vast majority of judge votes are in easy cases, where extrajudicial factors may play a smaller role. In hard cases, human biases could tip the swing vote, whose importance is magnified when examining case outcomes.

The continuous voice trait features improve case-wise accuracy by 2 percentage points[16] and decrease justice-wise accuracy by 0.6 percentage points over the 1998 to 2012 period where I had human ratings. Because my continuous trait rating prediction model does not have good predictive power, I do not predict continuous trait ratings or evaluate the model over the 1955–1997 and 2013–2014 period. Instead, I evaluate the model with the binary voice features, which improve case-wise accuracy by 1.1 percentage points and decreases justice-wise accuracy by 0.1 percentage points over the whole 1946–2014 period.

When I include both continuous voice and continuous image ratings (from 1998– 2012), I improve case-wise accuracy by 1 percentage point and improve the justice-wise accuracy by 0.8 percentage points. When I include binary voice and continuous image ratings (from 1980–2014), I improve case-wise accuracy by 0.9 percentage points and decrease justice-wise accuracy by 0.4 percentage points.

I used random forest to select features.[17] I observe that performance of the baseline model drops with the naïve addition of features, but after executing the random forest model again after feature selection, the predictive accuracy improves.

[16] Because there can be multiple lawyers on the petitioner or the respondent side, I average ratings for lawyers on each side.

[17] I initialized the number of features to 30, incremented the feature number by 10, and set a feature limit of 200. A brief schematic is as follows: 1) Fit the random forest classifier on the data, 2) Extract the feature

Justice-wise outcomes

Feature(s) added	Baseline	Feature Addition	Feature Addition and Selection
Image	0.645	0.640	0.667
Voice traits (continuous)	0.649	0.643	0.653
Voice traits (binary)	0.649	0.648	0.645
Image + Voice traits (continuous)	0.649	0.657	0.667
Image + Voice traits (binary)	0.639	0.635	0.665

Case-wise outcomes

Feature(s) added	Baseline	Feature Addition	Feature Addition and Selection
Image	0.638	0.656	0.688
Voice traits (continuous)	0.668	0.688	0.687
Voice traits (binary)	0.634	0.645	0.644
Image + Voice traits (continuous)	0.669	0.679	0.693
Image + Voice traits (binary)	0.636	0.647	0.667

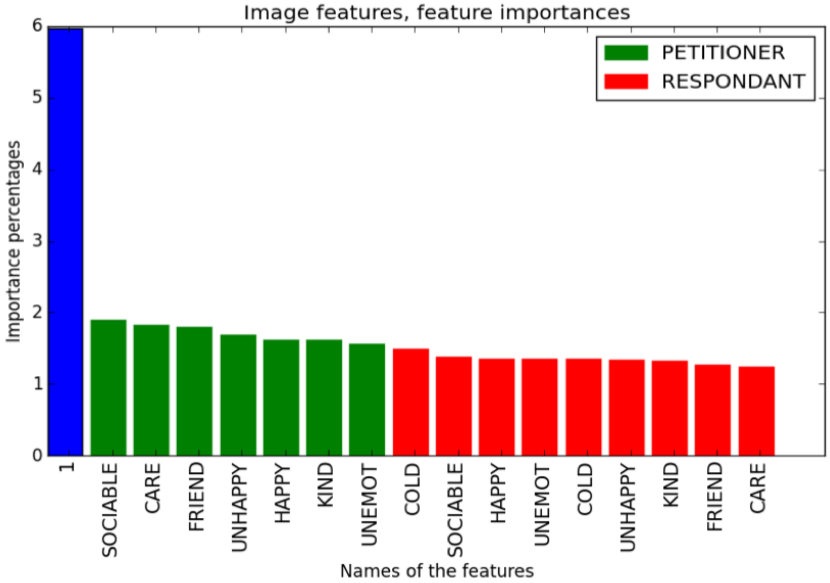
Feature Weights

The following charts show the feature weights for the image, audio, and image + audio features in their respective models. The blue bar with label "1" corresponds to the most

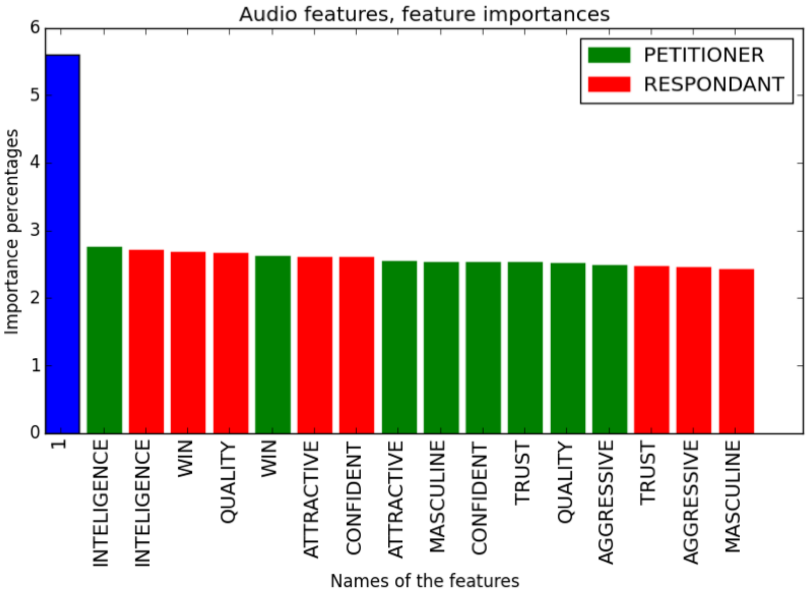
importance weights, 3) Select the top number of features (30), 4) Calculate the mean accuracy score, 5) Increment the number of features and repeat until I hit the limit of 200.

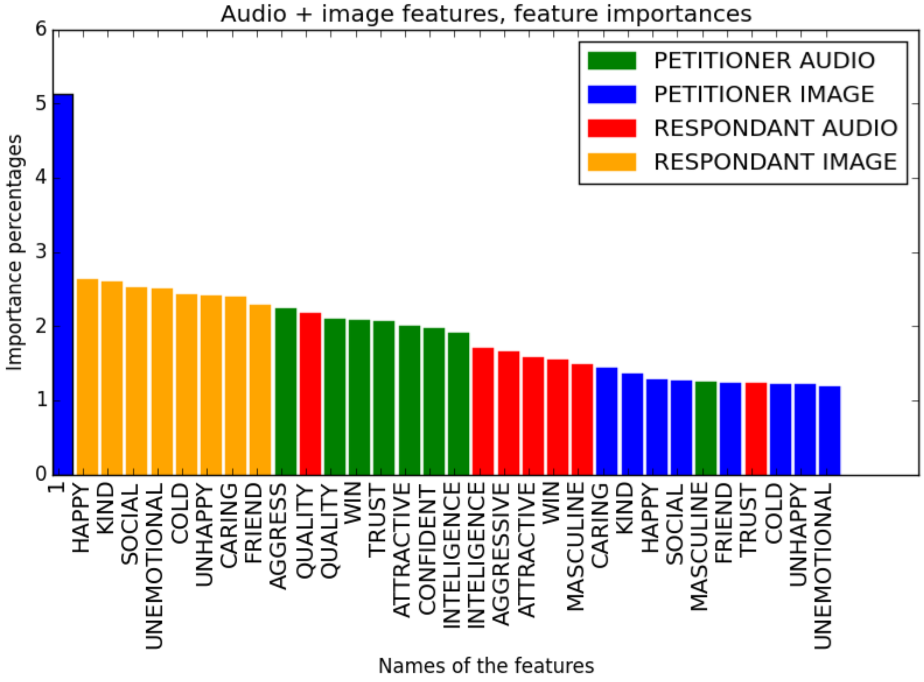
important feature present in the model.[18] Since this is a random forest model, the feature charts do not speak to the directionality of the features' effects. Thus, based only on this feature importance analysis I cannot say whether having a sociable face is associated with winning or losing.

Additionally, to address the question of whether my image features were really picking up lawyer gender, I ran my image feature model with a gender variable. I coded male lawyers as 1, and female lawyers as 0. I averaged the gender flags across lawyers on each side (petitioner and respondent). Where I have no lawyer images for a side, I used the average gender score (~0.89). I found that the average gender scores for petitioner and respondent sides were not in the top 30 most important features, and including the gender variables did not increase the accuracy of the model.



[18] In the "only image features", "image and audio features", and "only audio features (classification)" models, the most important feature (shown in the above charts as "1"), is "justice_previous_lc_direction_diff" (the difference between the lower court disposition direction and the justice's previous direction). For the "only audio features (regression)" model, the most important feature is "justice_cumulative_lc_direction_diff" (the difference between the lower court disposition direction and the justice's cumulative direction). Disposition direction is a measure of whether the decision of the court whose decision the Supreme Court reviewed was itself liberal or conservative. Previous refers to previous Supreme Court term and cumulative refers to all prior terms. As such, these two indicators are measurements related to ideology, and in particular, the ideological differences between the Justice and the lower court opinion.

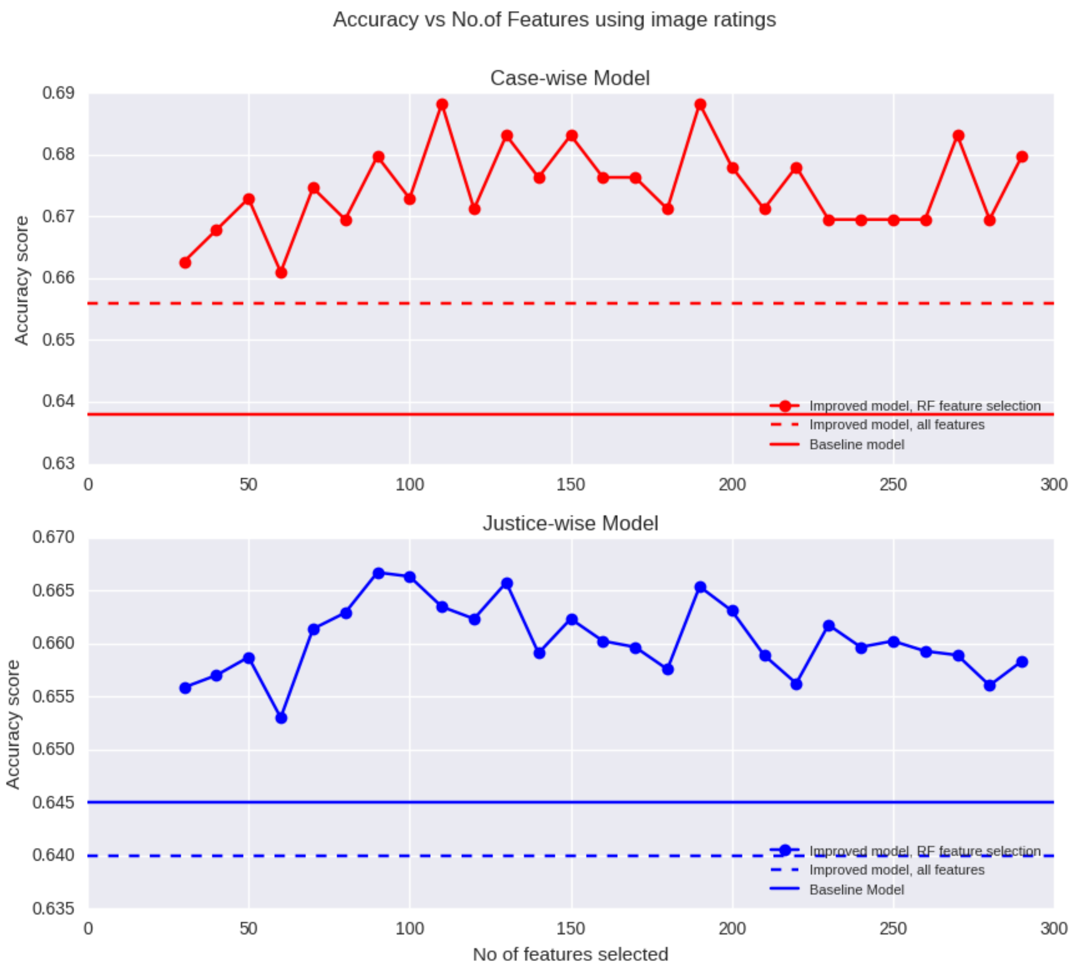

Audio features, feature importances
PETITIONER
RESPONDANT
Importance percentages
0
1
2
3
4
5
6
1
INTELIGENCE
INTELIGENCE
WIN
QUALITY
WIN
ATTRACTIVE
CONFIDENT
ATTRACTIVE
MASCULINE
CONFIDENT
TRUST
QUALITY
AGGRESSIVE
TRUST
AGGRESSIVE
MASCULINE
Names of the features


Audio + image features, feature importances
PETITIONER AUDIO
PETITIONER IMAGE
RESPONDANT AUDIO
RESPONDANT IMAGE
Importance percentages
0
1
2
3
4
5
6
1
HAPPY
KIND
SOCIAL
UNEMOTIONAL
COLD
UNHAPPY
CARING
FRIEND
AGGRESS
QUALITY
QUALITY
WIN
TRUST
ATTRACTIVE
CONFIDENT
INTELIGENCE
INTELIGENCE
AGGRESSIVE
ATTRACTIVE
WIN
MASCULINE
CARING
KIND
HAPPY
SOCIAL
MASCULINE
FRIEND
TRUST
COLD
UNHAPPY
UNEMOTIONAL
Names of the features

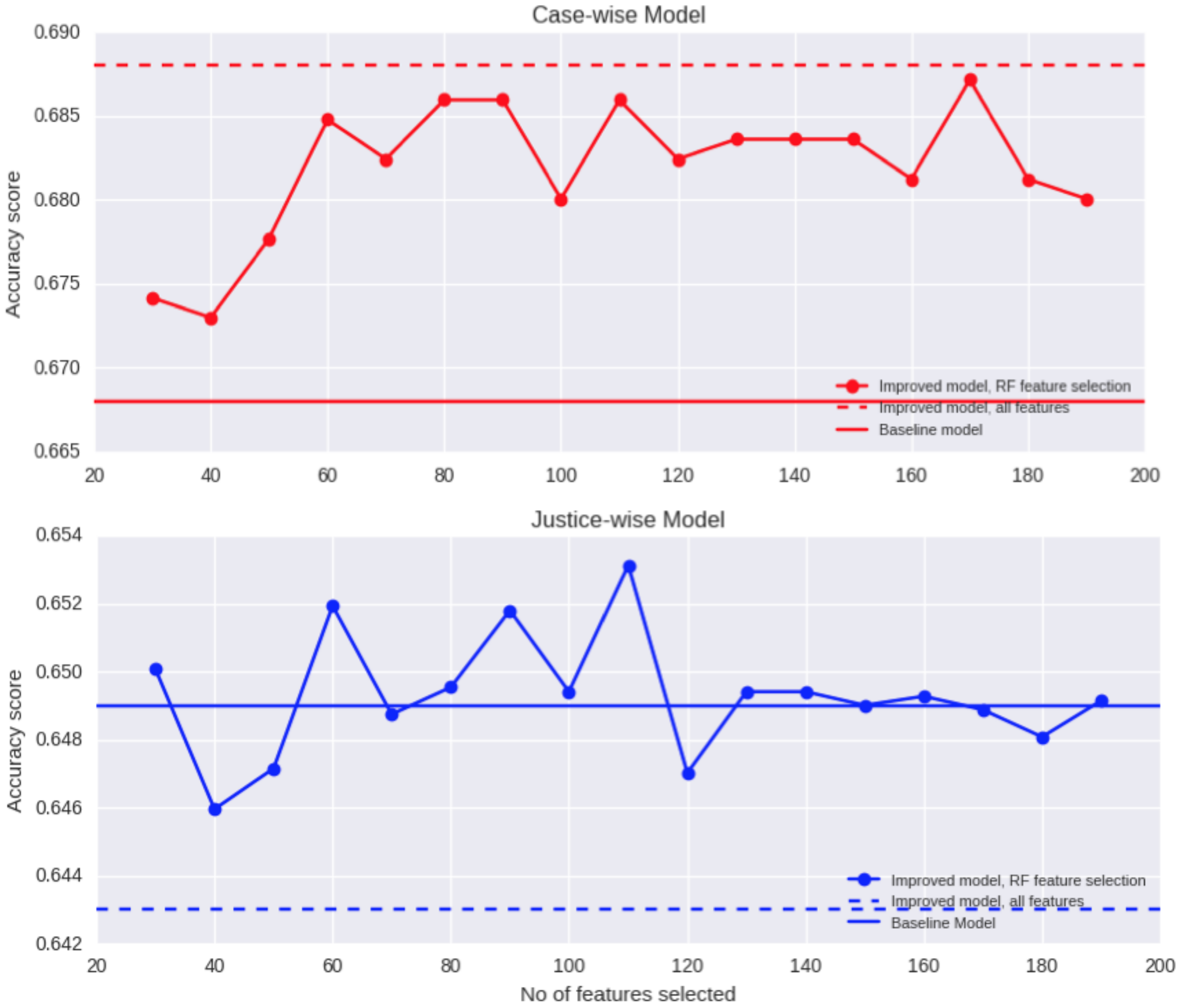
Performance

After adding my image/audio features I performed a feature selection routine. The following charts show the performance of the models varying the maximum number of features selected.

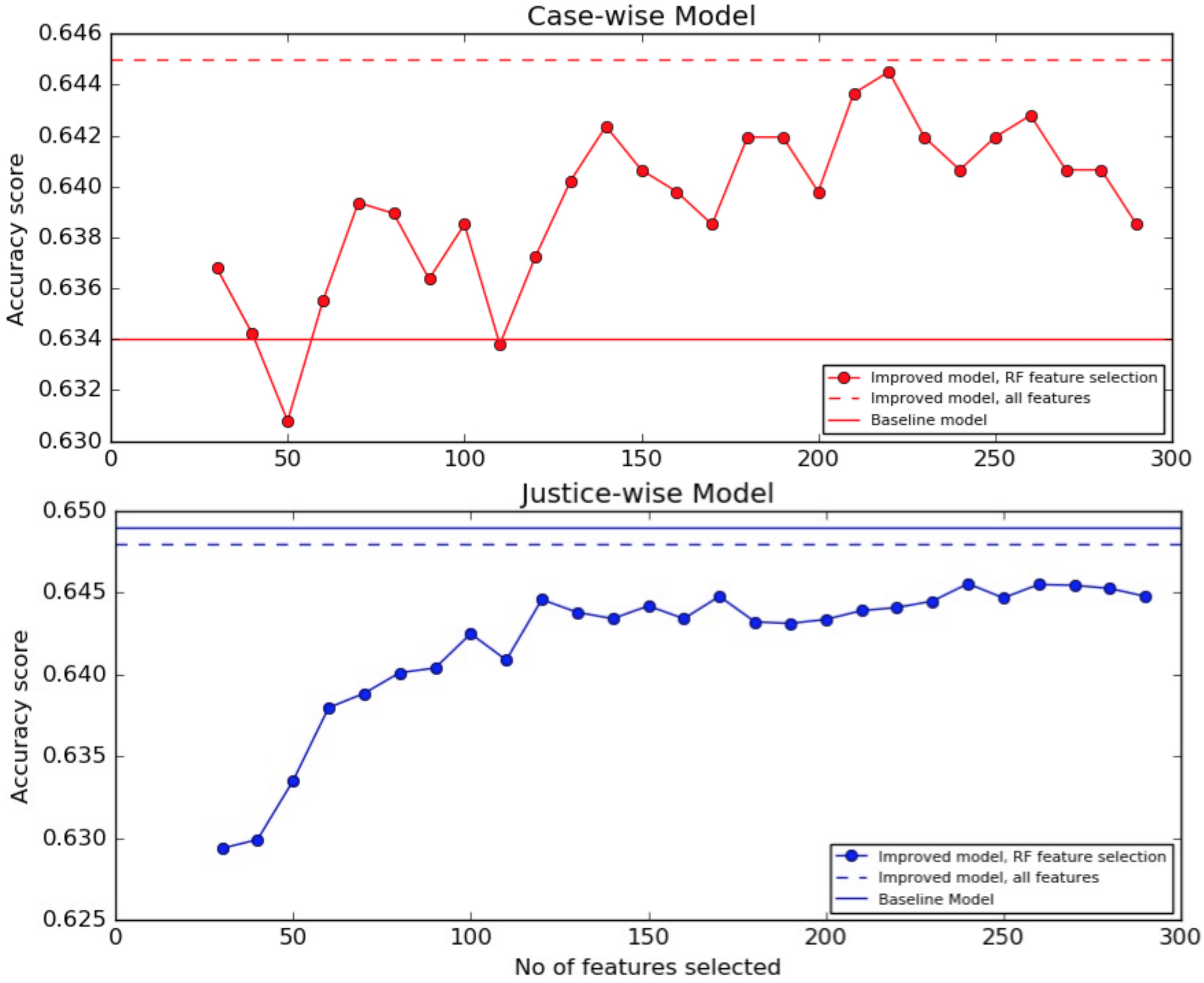
1. When only image features are considered:



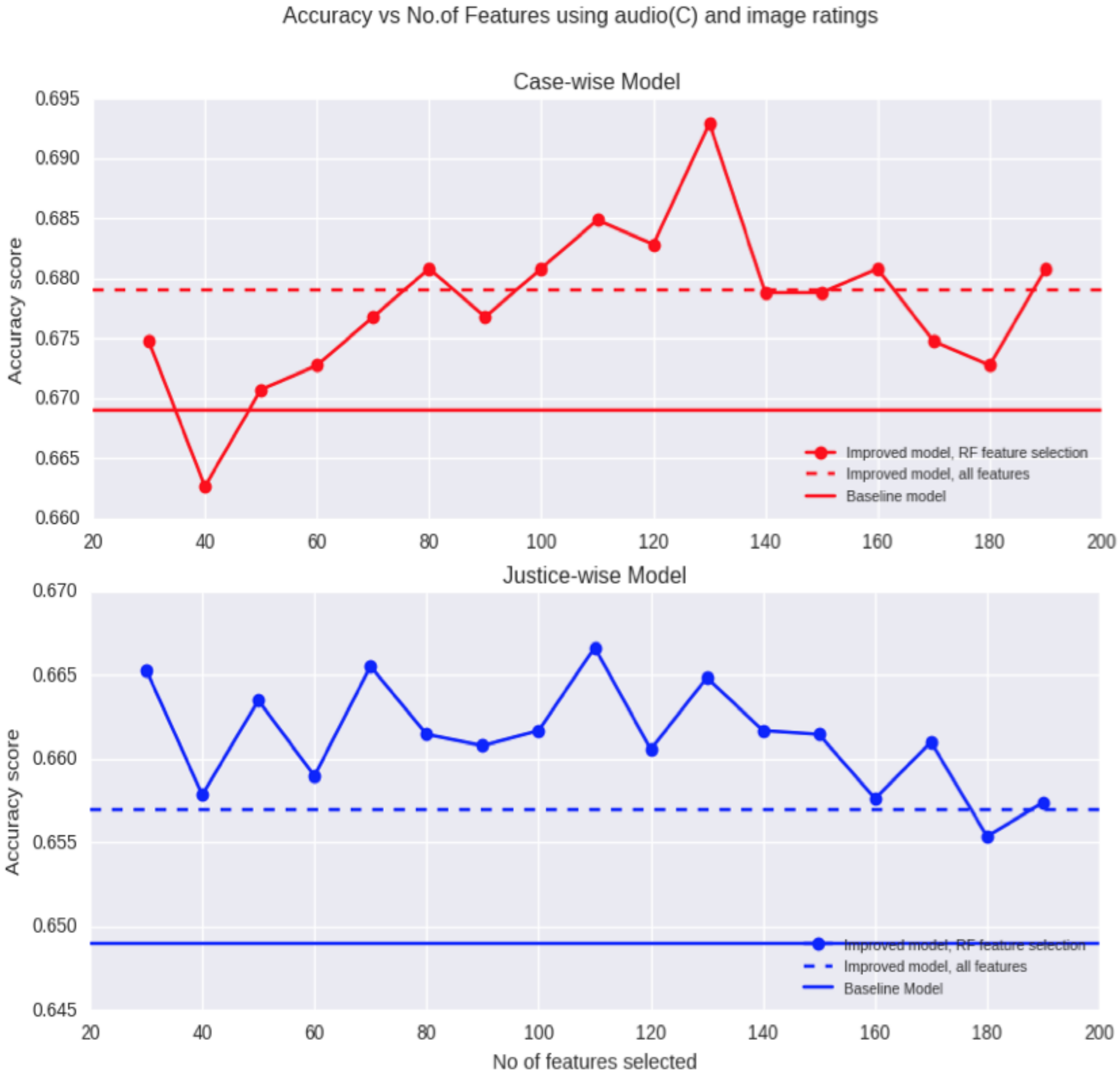
2. When only audio features (continuous ratings of voices from 1998–2012) are considered:



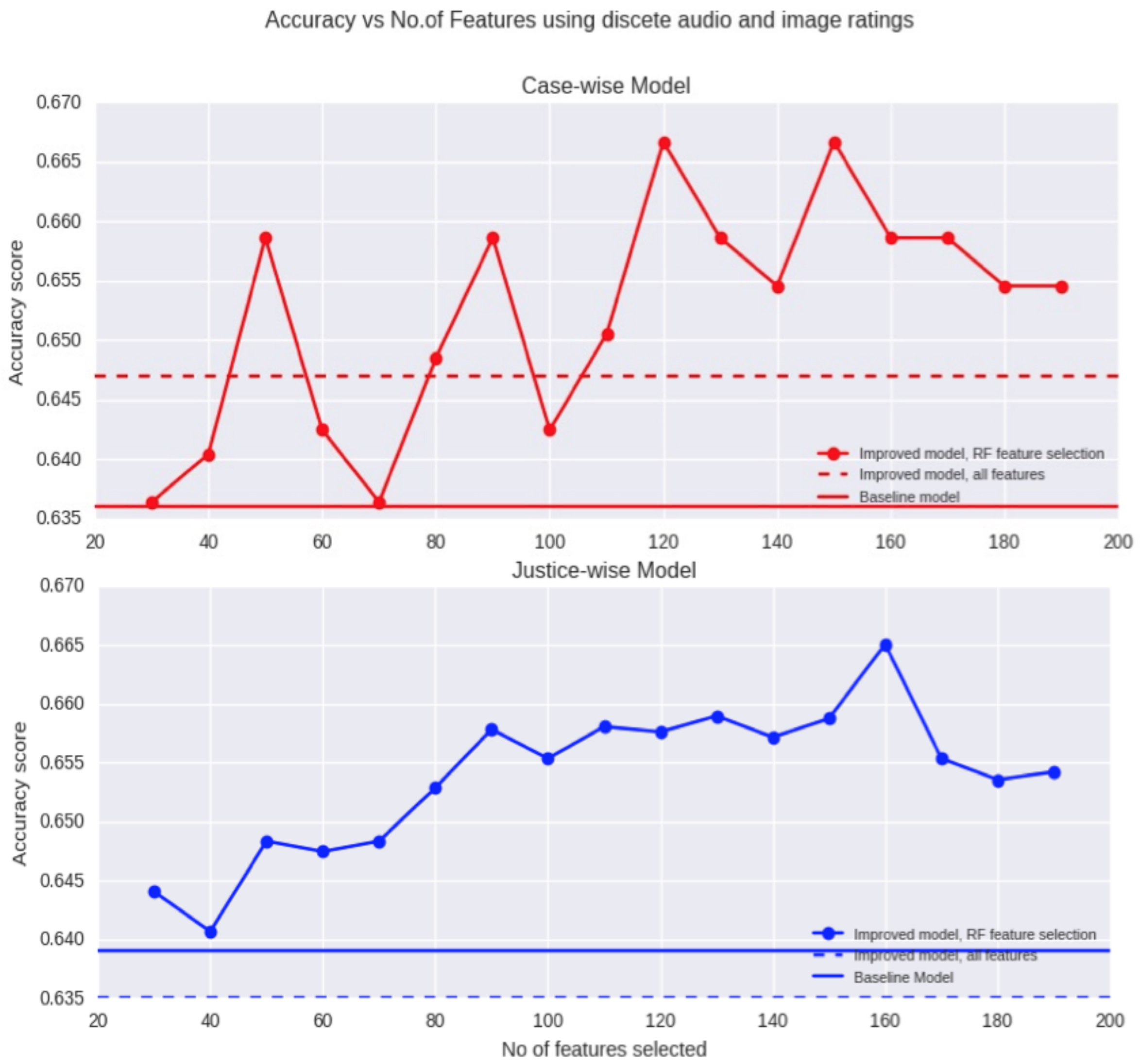
3. When only audio features (binary ratings of voices from 1946–2014) are considered:



4. When continuous voice ratings and image features are included (1998–2012):



5. When binary voice ratings and image features are included (1980–2014):



Discussion

Chen, et al. (2016, 2017) document in an econometric analysis that vocal features impact court outcomes. This paper shows that the best predictive model of Supreme Court votes improves with the addition of facial and voice characteristics of Supreme Court advocates. The improvement appears robust for predicting Supreme Court case outcomes and appears limited for predicting Supreme Court Justice votes, similar to the finding of Chen, et al. (2017). One interpretation of this difference is that hard or close cases may be more easily swayed by human biases. I also observed that due to the increase in the number of important features and decrease in available training data after incorporating image and audio features, a model that includes all the features can overfit, which I resolved by applying feature selection. A surprising finding is that these advocate characteristics received half as much in importance weight as the most important feature typically attributed to political ideology. However, an open question is whether the advocate characteristics are a signal of political ideology (Ash, et al. 2017).

An extension of my study can see if my predictive model improves when I count the number of times that Justices interrupt the advocate. I might also focus on the quantity or content of the interruption (features extracted from the text of the transcripts) or the quality of the interruption (features extracted from audio of the justice's interruption). I might also consider using a richer characterization of the audio clips rather than MFCC. Much richer audio characteristics of speech could be used to predict the trait with lesser error percentage. In ongoing work, one of the authors is collecting ratings for the 1946–1997 and 2013–2014 period and using these as inputs in a linear regression model, and these inputs may also be used in a prediction model. The lawyers' images might also be rated directly by humans rather than using a rater database to predict the traits.

References

Ash, Elliott, Daniel L. Chen and Suresh Naidu, “Ideas Have Consequences: The Impact of Law and Economics on American Justice,” working paper.

Ash, Elliott and Daniel L. Chen, “Dialects of Ideology,” working paper.

Berdejo, Carlos and Daniel L. Chen, “Electoral Cycles Among U.S. Courts of Appeals Judges.” The Journal of Law and Economics, forthcoming.

Bainbridge, Wilma A., Phillip Isola, and Aude Oliva. "The intrinsic memorability of face photographs." Journal of Experimental Psychology: General 142.4 (2013): 1323.

Carlos Berdejo and Daniel L. Chen, “Electoral Cycles Among U.S. Courts of Appeals Judges.” The Journal of Law and Economics, forthcoming.

Black, Ryan C., Sarah A. Treul, Timothy R. Johnson, and Jerry Goldman. "Emotions, Oral Arguments, and Supreme Court Decision Making." The Journal of Politics 73, no. 2 (2011): 572-581.

Chen, D. L. (2014), “Can markets overcome repugnance? Muslim trade response to anti-Muhammad Cartoons”, working paper, ETH Zurich, Mimeo.

Chen, D. L. (2015), “Can markets stimulate rights? on the alienability of legal claims”, RAND Journal of Economics, Vol. 46 No. 1, pp. 23–65.

Chen, D. L. (2005), “Gender violence and the price of virginity: theory and evidence of incomplete marriage contracts”, working paper, University of Chicago, Chicago, USA, Mimeo, November.

Chen, D. L. (2017), “Implicit egoism in sentencing decisions: first letter name effects with randomly assigned defendants”, working paper No. 16-726, Toulouse School of Economics, Toulouse, 31 May.

Chen, D. L. (2017), “Markets, morality, and economic growth: competition affects deontological judgment”, working paper, Toulouse School of Economics, Toulouse, France.

Chen, D. L. (2017), “Mood and the malleability of moral reasoning”, working paper, Toulouse School of Economics, Toulouse, France.

Chen, D. L. (2017), “Priming ideology: why presidential elections affect U.S. judges”, Journal of Law and Economics, Forthcoming.

Chen, D. L. (2017), “The deterrent effect of the death penalty? Evidence from British commutations during World War I, working paper, Toulouse School of Economics, Toulouse, France.

Chen, D. L., Cui, X., Shang, L. and Zheng, J. (2016), “What matters: agreement among U.S. courts of appeals judges”, Journal of Machine Learning Research, Forthcoming.

Chen, D. L., Dunn, M., Sagun L. and Sirin, H. (2017), “Early predictability of asylum court decisions”, in Proceedings of the Association for Computing Machinery Conference on Artificial Intelligence and the Law, Forthcoming.

Chen, D. L. and Eagel, J. (2016), “Can Machine Learning Help Predict the Outcome of Asylum Adjudications?” in Proceedings of the Association for Computing Machinery Conference on Artificial Intelligence and the Law, Forthcoming.

Chen, D. L., Halberstam, Y. and Yu, A. (2017), “Covering: mutable characteristics and perceptions of voice in the U.S. supreme court”, working paper, No.16-680, Toulouse School of Economics, Toulouse, France.

Chen, D. L., Halberstam, Y. and Yu, A. (2016), “Perceived masculinity predicts U.S. supreme court outcomes”, PLOS ONE, Vol. 11 No. 10, pp.1–20, e0164324.

Chen, D. L. and Horton J. (2016), “Are Online Labor Markets Spot Markets for Tasks? A Field Experiment on the Behavioral Response to Wage Cuts”, Information Systems Research, Vol. 27 No. 2, pp. 403–423.

Chen, D. L., Levonyan, V. and Yeh, S. (2017), “Policies Affect Preferences: Evidence from Random Variation in Abortion Jurisprudence”, working paper No. 16-723, Toulouse School of Economics, Toulouse, 7 March.

Chen, D. L. and Lind, J. (2017), “The Political Economy of Beliefs: Why Fiscal and Social Conservatives and Liberals Come Hand-in-Hand”, working paper No. 16-722, Toulouse School of Economics, Toulouse, 30 June.

Chen, D. L., and Loecher, M. (2017), “Events unrelated to crime predict criminal sentence length”, working paper, Toulouse School of Economics, Toulouse, France, 20 October.

Chen, D. L., Moskowitz, T. and Shue, K. (2016), “Decision making under the gambler’s fallacy: evidence from asylum judges, loan officers, and baseball umpires”, The Quarterly Journal of Economics, Vol. 131 No. 3, pp. 1181–1242.

Chen, D. L. and Schonger, M. (2017), “A theory of experiments: invariance of equilibrium to the strategy method of elicitation and implications for social preferences”, working paper No. 16-724, Toulouse School of Economics, Toulouse, 7 March.

Chen, D. L. and Schonger, M. (2017), “Social preferences or sacred values? theory and evidence of deontological motivations”, working paper No. 16-714, Toulouse School of Economics, Toulouse, 30 June.

Chen, D. L., Schonger, M. and Wickens, C. (2016), “oTree—An open-source platform for laboratory, online, and field experiments”, Journal of Behavioral and Experimental Finance, Vol. 9 No. 1, pp. 88 – 97.

Chen, D. L. and Sethi, J. (2017), “Insiders, outsiders, and involuntary unemployment: Sexual harassment exacerbates gender inequality”, working paper No. 16-687, Toulouse School of Economics, Toulouse, 31 May.

Chen, D. L. and Yeh, S. (2014), “The construction of morals”, Journal of Economic Behavior and Organization, Vol. 104, pp. 84–105.

Chen, D. L. and Yeh, S. (2017), “How do rights revolutions occur? free speech and the first amendment”, working paper No. 16-705, Toulouse School of Economics, Toulouse, 7 March.

Epstein, Lee, William M. Landes, and Richard A. Posner. "Inferring the winning party in the Supreme Court from the pattern of questioning at oral argument." The Journal of Legal Studies 39.2 (2010): 433-467.

Katz, Daniel Martin, Michael James Bommarito, and Josh Blackman. "Predicting the behavior of the supreme court of the united states: A general approach." Available at SSRN 2463244 (2014).

Lyons, James. "Python_speech_features." GitHub. N.p., n.d. Web. 14 May 2016.

Oyez. IIT Chicago-Kent College of Law, n.d. Web. 14 May 2016.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." Journal of Machine Learning Research 12, no. Oct (2011): 2825-2830.

Rojas, Mario, David Masip, Alexander Todorov, and Jordi Vitria. "Automatic prediction of facial trait judgments: Appearance vs. structural models." PloS one 6, no. 8 (2011): e23323.

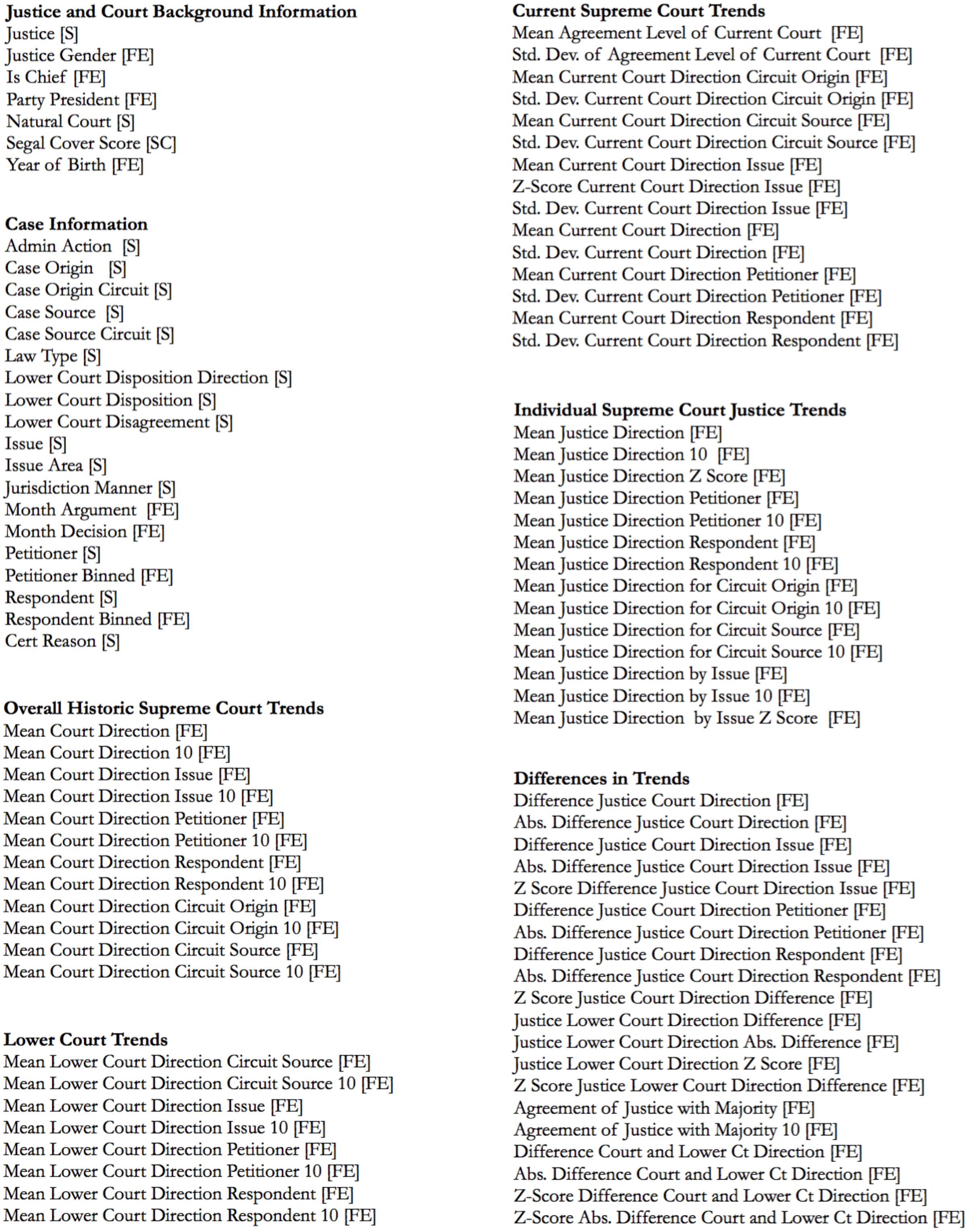
Rosenberg, Gerald N. The Hollow Hope: Can Courts Bring About Social Change? University of Chicago Press, 2008.

Schubert, James N., Steven A. Peterson, Glendon Schubert, and Stephen Wasby. "Observing Supreme Court oral argument: A biosocial approach."Politics and the Life Sciences (1992): 35-51.

Todorov, Alexander, Anesu N. Mandisodza, Amir Goren, and Crystal C. Hall. "Inferences of competence from faces predict election outcomes." Science 308, no. 5728 (2005): 1623-1626.

Appendix 1

List of variables employed in the baseline model (Katz et al.):



Justice and Court Background Information

Justice [S]
Justice Gender [FE]
Is Chief [FE]
Party President [FE]
Natural Court [S]
Segal Cover Score [SC]
Year of Birth [FE]

Case Information

Admin Action [S]
Case Origin [S]
Case Origin Circuit [S]
Case Source [S]
Case Source Circuit [S]
Law Type [S]
Lower Court Disposition Direction [S]
Lower Court Disposition [S]
Lower Court Disagreement [S]
Issue [S]
Issue Area [S]
Jurisdiction Manner [S]
Month Argument [FE]
Month Decision [FE]
Petitioner [S]
Petitioner Binned [FE]
Respondent [S]
Respondent Binned [FE]
Cert Reason [S]

Overall Historic Supreme Court Trends

Mean Court Direction [FE]
Mean Court Direction 10 [FE]
Mean Court Direction Issue [FE]
Mean Court Direction Issue 10 [FE]
Mean Court Direction Petitioner [FE]
Mean Court Direction Petitioner 10 [FE]
Mean Court Direction Respondent [FE]
Mean Court Direction Respondent 10 [FE]
Mean Court Direction Circuit Origin [FE]
Mean Court Direction Circuit Origin 10 [FE]
Mean Court Direction Circuit Source [FE]
Mean Court Direction Circuit Source 10 [FE]

Lower Court Trends

Mean Lower Court Direction Circuit Source [FE]
Mean Lower Court Direction Circuit Source 10 [FE]
Mean Lower Court Direction Issue [FE]
Mean Lower Court Direction Issue 10 [FE]
Mean Lower Court Direction Petitioner [FE]
Mean Lower Court Direction Petitioner 10 [FE]
Mean Lower Court Direction Respondent [FE]
Mean Lower Court Direction Respondent 10 [FE]

Current Supreme Court Trends

Mean Agreement Level of Current Court [FE]
Std. Dev. of Agreement Level of Current Court [FE]
Mean Current Court Direction Circuit Origin [FE]
Std. Dev. Current Court Direction Circuit Origin [FE]
Mean Current Court Direction Circuit Source [FE]
Std. Dev. Current Court Direction Circuit Source [FE]
Mean Current Court Direction Issue [FE]
Z-Score Current Court Direction Issue [FE]
Std. Dev. Current Court Direction Issue [FE]
Mean Current Court Direction [FE]
Std. Dev. Current Court Direction [FE]
Mean Current Court Direction Petitioner [FE]
Std. Dev. Current Court Direction Petitioner [FE]
Mean Current Court Direction Respondent [FE]
Std. Dev. Current Court Direction Respondent [FE]

Individual Supreme Court Justice Trends

Mean Justice Direction [FE]
Mean Justice Direction 10 [FE]
Mean Justice Direction Z Score [FE]
Mean Justice Direction Petitioner [FE]
Mean Justice Direction Petitioner 10 [FE]
Mean Justice Direction Respondent [FE]
Mean Justice Direction Respondent 10 [FE]
Mean Justice Direction for Circuit Origin [FE]
Mean Justice Direction for Circuit Origin 10 [FE]
Mean Justice Direction for Circuit Source [FE]
Mean Justice Direction for Circuit Source 10 [FE]
Mean Justice Direction by Issue [FE]
Mean Justice Direction by Issue 10 [FE]
Mean Justice Direction by Issue Z Score [FE]

Differences in Trends

Difference Justice Court Direction [FE]
Abs. Difference Justice Court Direction [FE]
Difference Justice Court Direction Issue [FE]
Abs. Difference Justice Court Direction Issue [FE]
Z Score Difference Justice Court Direction Issue [FE]
Difference Justice Court Direction Petitioner [FE]
Abs. Difference Justice Court Direction Petitioner [FE]
Difference Justice Court Direction Respondent [FE]
Abs. Difference Justice Court Direction Respondent [FE]
Z Score Justice Court Direction Difference [FE]
Justice Lower Court Direction Difference [FE]
Justice Lower Court Direction Abs. Difference [FE]
Justice Lower Court Direction Z Score [FE]
Z Score Justice Lower Court Direction Difference [FE]
Agreement of Justice with Majority [FE]
Agreement of Justice with Majority 10 [FE]
Difference Court and Lower Ct Direction [FE]
Abs. Difference Court and Lower Ct Direction [FE]
Z-Score Difference Court and Lower Ct Direction [FE]
Z-Score Abs. Difference Court and Lower Ct Direction [FE]